

CLAIMS

1 1. A computer-implemented method for retrieving documents comprising:
2 inputting the text of one or more documents, wherein each document includes
3 human readable words;
4 creating context windows around each said word in each document;
5 generating a statistical evaluation of the characteristics of all of the windows,
6 wherein the results are not a function of the order of the appearance of words within
7 each window; and
8 combining the results of the statistical evaluation for each window.

9 2. The method according to Claim 1 further comprising:
10 determining the likelihood of documents having predetermined characteristics based on the
11 combined statistical evaluation for each window.

12 3. The method according to Claim 2 further comprising:
13 assigning a document identifier to each document and context window position; and
14 determining the document identifier of at least one document having said predetermined
15 characteristics.

16 4. The method according to Claim 1 further comprising:
17 defining a plurality of document categories; and

3 determining the category of a particular document based on the combined statistical evaluation
4 for each window.

1 5. The method according to Claim 1 further comprising:
2 determining the word that is in the center of a particular window based on the combined
3 statistical evaluation for each window.

1 6. The method according to Claim 1 wherein the step of generating a statistical evaluation
2 further includes counting the occurrences of particular words and particular documents and tabulating
3 totals of the counts.

1 7. The method according to Claim 6 wherein the step of generating a statistical evaluation
2 further includes the step of generating counts about singular word occurrences and about pair-wise
3 occurrences.

1 8. The method according to Claim 7 further comprising the step of pruning the number
2 of pair-wise counts.

1 9. The method according to Claim 8 wherein the step of pruning further includes the
2 steps of monitoring the amount of memory used for the pair-wise counts and pruning when a
3 predetermined threshold of memory has been exceeded for the pair-wise counts.

1 10. The method according to Claim 6 wherein the step of generating a statistical evaluation
2 further includes the step of determining probabilities of particular words appearing in particular
3 documents based on the counts.

1 11. The method according to Claim 10 wherein the step of generating a statistical
2 evaluation further includes determining conditional probabilities of particular words appearing in
3 particular documents based on the counts.

1 12. The method according to Claim 11 further comprising the step of calculating a
2 conditional probability based on a Simple Bayes statistical model.

1 13. The method according to Claim 1 wherein the step of creating context windows around
2 each word further comprises the step of selecting the words appearing before and after each word
3 by a predetermined amount in the document and including those selected words in the window.

1 14. The method according to Claim 13 wherein the word around which each window is
2 created is not included in the window.

1 15. The method according to Claim 1 further comprising normalizing the combined results
2 of the statistical evaluation for the windows.

1 16. The method according to Claim 1 wherein the step of evaluating further comprises,
2 determining a measure of mutual information.

1 17. The method according to Claim 1 wherein the step of combining includes averaging
2 probability assessments.

1 18. A computer system comprising:
2 storage unit for receiving and storing a plurality of documents, wherein each
3 document includes human readable words; means for creating context windows around
4 each said word in each document;
5 means for generating a statistical evaluation of the content of each window,
6 wherein the order of the appearance of words within each window is not used in the
7 statistical evaluation;
8 means for combining the results of the statistical evaluation for each window;
9 and
10 means for determining the probabilities of documents having predetermined
11 characteristics based on the combined statistical evaluation for each window.

1 19. The computer system according to Claim 18 further comprising:
2 a document identifier assigned to each document; and
 means for determining the document identifier of at least one document having said
predetermined characteristics.

1 20. The computer system according to Claim 18 further comprising:
2 a plurality of document categories; and
3 means for determining the category of a particular document based on the combined statistical
4 evaluation for each window.

1 21. The computer system according to Claim 18 further comprising:
2 means for determining the word that is in the center of a particular window based on the
3 combined statistical evaluation for each window.

1 22. The computer system according to Claim 18 wherein the step of generating a statistical
2 evaluation further includes counting the occurrences of particular words and particular documents and
3 tabulating totals of the counts.

1 23. The computer system according to Claim 22 wherein the means for generating a
2 statistical evaluation further includes means for determining probabilities of particular words
3 appearing in particular documents based on the counts.

1 24. The computer system according to Claim 23 wherein the means for generating a
2 statistical evaluation further includes means for determining conditional probabilities of particular
3 words appearing in particular documents based on the counts.

•

1 25. The computer system according to Claim 18 wherein the means for creating context
2 windows around each word further comprises means for selecting the words appearing before and
3 after each word by a predetermined amount in the document and including those selected words in
4 the window.

1 26. A computer program product comprising:

2 a computer program storage device;

3 computer-readable instructions on the storage device for causing a computer
4 to undertake method acts to facilitate retrieving documents, the method acts
5 comprising:

6 inputting the text of one or more documents, wherein each document includes
7 human readable words;

8 creating context windows around each said word in each document;

9 generating a statistical evaluation of the characteristics of each window,
10 wherein the results are not a function of the order of the appearance of words within
11 each window; and

12 combining the results of the statistical evaluation for each window.

1 27. The computer program product according to Claim 26 further comprising:

2 determining the likelihood of documents having predetermined characteristics based on the
3 combined statistical evaluation for each window.